Erschienen in: Haß, Ulrike (Hrsg.): Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. - Berlin/New York: de Gruyter, 2005. S. 71-90.
(Schriften des Instituts für Deutsche Sprache 12)

*Ulrich Schnörch*

# Die *elexiko*-Stichwortliste



## 1. Die vollständige Neuerarbeitung einer Stichwortliste für *elexiko*

### 1.1 Grundsätzliches und Vorüberlegungen

Eine Stichwortliste stellt prinzipiell die Basis, das Gerüst eines Wörterbuches und natürlich auch eines lexikologisch-lexikografischen Informationssystems wie *elexiko* dar. Je nachdem, ob man beispielsweise im Rahmen der einsprachigen Printlexikografie ein einbändiges, allgemeinsprachliches Wörterbuch plant, gar ein mehrbändiges, eines, das ausschließlich Fremdwörter enthalten soll oder nur den sog. Grundwortschatz – die konzeptionellen Vorüberlegungen beeinflussen Umfang, Art und Ausgewogenheit der aufzunehmenden Wörter, also die Stichwortliste maßgeblich.

Die Makrostruktur in der Printlexikografie (vgl. Wiegand 1989) setzt die Erstellung einer möglichst kompletten Stichwortliste als Grundlage der eigentlichen Wortartikelproduktion voraus, da hier die alphabetische Anordnung der Stichwörter im gesamten Wörterbuch und darüber hinaus innerhalb der Einzelbuchstaben ein Voranschreiten bei der Wortartikelproduktion von A-Z i. d. R. erforderlich macht. Die LexikografInnen müssen möglichst von Anfang an die Reihen- bzw. Abfolge der Stichwörter kennen und diese auch einhalten; der einmal gedruckte Band lässt nachträgliche Ergänzungen oder Streichungen nicht mehr zu – allenfalls in einer Neuauflage. Aus diesem Grund ist die Stichwortliste – sobald sie einmal erstellt ist – weitestgehend unveränderbar und starr.

Die elektronische Publikationsform von *elexiko* lässt demgegenüber ein weitaus dynamischeres Stichwortlistenkonzept zu: Möglich wird dies v. a.,

weil das Alphabet als ordnende, den Arbeitsprozess (und später die Such- und Findestrategien) maßgeblich steuernde Instanz keine Rolle spielt; es können im elektronischen Informationssystem andere Bearbeitungsmechanismen (und Recherchetechniken) angewandt werden.

Auf der Basis einer kompletten Stichwortliste lassen sich einzelne Wortschatzbereiche nach unterschiedlichen (z. B. inhaltlichen) Gesichtspunkten auswählen; diese sog. (Wortschatz-) Module können von LexikografInnen bearbeitet werden, ohne dass sie dabei im Geringsten auf eine alphabetische Ordnung achten müssen; im Gegensatz zu Printwörterbüchern sind darüber hinaus nachträgliche Ergänzungen und Streichungen im Rahmen der Makrostruktur ohne Probleme möglich.

Konsequent weitergedacht würde die modular angelegte Konzeption von *elexiko* unter Umständen die Voraberstellung einer sehr umfangreichen Stichwortliste gar nicht zwingend erforderlich machen. Die Wortartikelproduktion lediglich im Rahmen bestimmter kleinerer, inhaltlich bzw. funktional definierter Wortschatzausschnitte, gleichsam ein organisches Anwachsen des Stichwortbestandes wäre sehr wohl denkbar.

Doch selbst bei einem modularen, dynamischen Konzept sprechen zumindest folgende Überlegungen dafür, mit der Erstellung einer weitgehend kompletten und umfangreichen Stichwortliste zu beginnen:

1. Um qualitativ eine homogene Liste erstellen zu können, muss man sich im Klaren sein, was man überhaupt unter einem Stichwort versteht. Es ist also nötig, eine funktional und formal an den Erfordernissen ausgerichtete, für das Projekt möglichst allgemein verbindliche Stichwortkonzeption zugrunde zu legen.
2. Eine auch noch so umfassend erscheinende Stichwortliste stellt quantitativ bereits eine konzeptionell begründete **Auswahl** aus dem gegen unendlich gehenden Wortbestand einer Sprache dar:
   Sie bildet (wie oben skizziert) das methodisch auf das Ziel des Wörterbuchprojekts abgestimmte Fundament für die Wortartikelproduktion – und *elexiko* hat das Ziel, die Beschreibung eines sehr großen Wortschatzbestandes der geschriebenen deutschen (Gegenwarts-) Sprache in Angriff zu nehmen: ca. 300.000 Stichwörter. Zukünftig können die LexikografInnen aus diesem großen, einmal systematisch erarbeiteten Wortschatzpool je nach Bedarf speziellere (und nicht vom Alphabet geleitete) Auswahlen treffen;[1] sie können nachträglich

[1] Eine solche Auswahl stellt beispielsweise der sog. *elexiko*-'Demonstrationswortschatz' dar, mit dem alle Arbeitsschritte der Artikelproduktion (mit Ausnahme der Angaben zur Wortbildungsproduktivität und zur Diachronie) ausgiebig getestet wurden, vgl. www. http://www.elexiko.de/).

neue Stichwörter ergänzen oder Streichungen vornehmen, freilich im Rahmen der festgelegten Richtlinien bzw. innerhalb eines bestimmten Toleranzspielraumes (der allerdings deutlich größer ist als bei Printwörterbüchern).

3. In *elexiko* sind zwei Arten des Informationsauf- bzw. -ausbaus geplant: ein ‚vertikaler', d. h. in die Detailtiefe gehender bei der Produktion von einzelnen Wortartikeln, wie er von Printwörterbüchern her bekannt ist; die elektronische Publikation ermöglicht darüber hinaus auch einen ‚horizontalen' Informationsausbau, d. h. die (teil-)automatische Ergänzung bestimmter Angaben z. B. zur Orthografie, zu morphologischen Varianten u. a. über den gesamten Stichwortbestand hinweg; dies ist natürlich nur dann sinnvoll und überhaupt durchführbar, wenn die Stichwortliste komplett vorliegt.[2]
4. Um die Möglichkeiten des Hypertextes zu nutzen, ist ein Verlinkungssystem nötig. In Grundzügen ist dies mit den herkömmlichen Wörterbuchverweisen vergleichbar. Erst eine komplett angelegte Stichwortliste ermöglicht es, von Wortartikel a (ggf. Lesart x) auf Wortartikel b (ggf. Lesart y) zu verweisen. Die Stichwortliste bildet in diesem Kontext also gewissermaßen ein ‚Adressverzeichnis', den Bezugspunkt für ein geplantes, weit komplexeres Verlinkungssystem.

## 1.2 Methodik

Grundsätzlich gibt es zwei unterscheidbare Ansätze, eine Stichwortliste zu erstellen:[3]

- Man kann bereits vorhandene Stichwortlisten auswählen, auswerten und schließlich mehr oder minder modifiziert übernehmen.
- Man kann sie auf der Basis von Korpora komplett neu erstellen.

In der Regel werden beide Methoden kombiniert, wobei man in Abhängigkeit zur Zielsetzung des lexikografischen Projekts den Schwerpunkt auf eine der

---

[2] Analog dazu sind zwei Arten von Veröffentlichungsschritten zu unterscheiden: Der modulweise (vergleichbar den [Teil-]Buchstabenlieferungen mehrbändiger Wörterbücher) sowie der kontinuierliche, d. h. der etappenweise Zuwachs bestimmter Dateneinheiten zum kompletten Wortbestand.

[3] Bei den folgenden Ausführungen geht es um die Erstellung einer Stichwortliste, die sich ausschließlich aus Einworteinheiten zusammensetzt. Die *elexiko*-Stichwortliste soll später jedoch um Mehrworteinheiten und Wortbildungsbestandteile ergänzt werden. Hierfür sind jedoch andere Methoden notwendig als die hier beschriebenen.

beiden Vorgehensweisen legen, die andere als Ergänzung bzw. Korrektiv einsetzen wird. Für die *elexiko*-Stichwortliste wurde im Wesentlichen die zweitgenannte Methode gewählt, und diese Entscheidung liegt nahe, denn *elexiko* hat sich dem Prinzip der Korpusbasiertheit verpflichtet: Ziel ist es, den Sprachgebrauch in der deutschen Schriftsprache seit 1946 möglichst umfassend zu beschreiben, und da es vom lexikografischen Standpunkt aus unmöglich ist, das Deutsche in seiner Gesamtheit zu dokumentieren, muss man sich auf einen fassbaren Wortschatzausschnitt beschränken; diese Voraussetzungen erfüllt das eigens zusammengestellte *elexiko*-Korpus (siehe dazu P. Storjohann, Das *elexiko*-Korpus, in diesem Band). Die Korpusbasiertheit als methodischer Grundsatz sollte konsequenterweise nicht nur für die Wortartikelproduktion gelten, sondern bereits für die Erarbeitung der zugrunde gelegten Stichwortliste maßgebend sein, auf den Punkt gebracht: *elexiko* heißt Lexikografie ab ovo.

Eine entscheidende Voraussetzung für die Wortartikelproduktion ist daher, dass in *elexiko* ausschließlich solche Stichwörter angesetzt werden, für die es tatsächlich Belege im *elexiko*-Korpus gibt: Es hätte insofern auch wenig Sinn gemacht, ausschließlich externe Stichwortlisten heranzuziehen und erst bei der Wortartikelproduktion zu überprüfen, ob die daraus übernommenen Stichwörter überhaupt im *elexiko*-Korpus vertreten sind, also eine korpusbasierte, lexikologisch-lexikografisch adäquate Beschreibung möglich ist.

Zusammengefasst bietet dieses Vorgehen zwei entscheidende Vorteile: Es werden keine Wörterbuchleichen aus anderen Listen übernommen, stattdessen können jedoch Lücken entdeckt und gefüllt werden. Dafür müssen aus einer nahezu unüberschaubaren Menge von ‚Wörtern', wie sie in Texten des Korpus vorkommen, über verschiedene, nachfolgend näher beschriebene Arbeitsschritte Stichwörter abgeleitet werden, wie sie der Konzeption von *elexiko* entsprechen.

Ausblickend ist die Entscheidung für die konzeptionelle Eigenständigkeit zugleich eine Voraussetzung für Aktualisierungen auf der Basis ständiger Neuakquisitionen im Bereich des *elexiko*-Korpus: Aktualisierungsroutinen liegen nämlich auf einer Linie mit dem eigenständigen, dynamischen Stichwortlistenkonzept.

## 2. Die Vorgehensweise

Gemäß der Zielvorgaben sollte für *elexiko* eine Stichwortliste in der Größenordnung von ca. 300.000 Einheiten erstellt werden. Nachfolgend werden die Arbeitsschritte im Einzelnen skizziert, die für die Umsetzung des Plans notwendig waren.

Beim Verfahren zur Erstellung der *elexiko*-Stichwortliste lassen sich zwei größere Teilschritte unterscheiden:

1. Der erste Teilschritt ist überwiegend computergestützt, automatisch abgelaufen: Der Schwerpunkt lag bei der Auswertung von Korpusmassendaten zum Zwecke der Generierung einer vorläufigen sog. Stichwortkandidatenliste.
2. Der zweite Teilschritt war weitgehend ‚manuell' zu bewältigen: Der Schwerpunkt liegt bei der redaktionellen lexikologisch-lexikografischen Überprüfung und Aufbereitung dieser Daten zur *elexiko*-Stichwortliste.

## 2.1 Der computergestützte, automatische Teilschritt

Für den ersten Teilschritt zeigte sich die AG-Korpustechnologie des IDS, Mannheim verantwortlich. Ihr Verdienst ist die Bereitstellung einer Liste von ca. 320.000 ‚Wörtern', die im Folgenden als Stichwortkandidaten bezeichnet werden, denn es sind vorläufige, d. h. bis zu einer weiteren Überprüfung im zweiten Teilschritt noch potentielle *elexiko*-Stichwörter. Bei der überwiegenden Mehrzahl der Fälle handelt es sich bereits um automatisch auf Grundformen (auch Nenn-, Zitier- bzw. kanonische Formen genannt) zurückgeführte ‚Wörter', also im Idealfall ‚Stichwörter', wie sie meist durch Fettdruck in Printwörterbüchern gekennzeichnet werden. Auf die technischen Details zur Ermittlung der 320.000 Stichwortkandidaten aus den Korpora wird nicht näher eingegangen. Stark vereinfachend und der eigentlichen Komplexität des Vorgangs nicht Rechnung tragend lässt sich zusammenfassend sagen, dass die Stichwortkandidaten durch die Kombination zweier Verfahren ermittelt wurden:

1. Mithilfe eines automatischen Lemmatisierers[4] wurden die Flexions- bzw. Paradigmenformen von Wörtern aus den Texten der IDS-Korpora geschriebener Gegenwartssprache[5] Wortformen zugewiesen, die gemäß der Annahmen des automatischen Lemmatisierers als ‚wörterbuchübliche Grund- oder Nennformen' gedeutet werden können, z. B. dem Infinitiv bei Verben, dem Nominativ, Singular bei Nomen, dem Positiv bei Adjektiven usw.; das sind die sog. Stichwortkandidaten. Zusätzlich wurden auch lexikografisch relevante Flexionsformen ei-

---

4 Nähere Einzelheiten dazu siehe: http://www.ids-mannheim.de/kt/dokumente/glemmrep.pdf.

5 Nähere Einzelheiten dazu siehe: http://www.ids-mannheim.de/kt/projekte/korpora/.

gens erfasst, beispielsweise Partizipien, unregelmäßige Pluralbildungen, die erste Person Präteritum der starken Verben u. Ä. Die Stichworthypothesen des automatischen Lemmatisierers können freilich von denen der Lexikografen abweichen. Dass es bei der automatischen Zuordnung von Flexions- und Grundformen in Einzelfällen zu Fehlern kommen kann, braucht nicht näher ausgeführt zu werden; Polysemie und Polyfunktionalität der sprachlichen Einheiten sind nur zwei Gründe: Handelt es sich beispielsweise bei großgeschriebenem *All* um *All* in der Lesart ‚Universum', oder um *all* in der Verwendung als großgeschriebenes Personalpronomen am Satzanfang? Ohne weitere Zusatzinformation bei der entsprechenden Wortform im Korpustext (Annotation) ist dem automatischen Lemmatisierer eine eindeutige Zuordnung in solchen und ähnlichen Fällen nicht möglich.

2. Um von Vornherein die Zahl insbesondere formal fehlerhaft angesetzter Stichwortkandidaten einzugrenzen, erfolgte ein sichernder Korrekturabgleich der Stichwortkandidatenliste mit Listen anderer Wörterbücher.

Die Stichwortkandidatenliste ist das Ergebnis eines Zusammenspiels der beiden Methoden ‚Korpusextraktion mit automatischer (Grund-) Formzusammenführung' und ‚korrektiver formaler Abgleich mit vorhandenen Wörterbuchlisten' im Anschluss daran. Das Resultat wäre allerdings noch um ein Vielfaches umfangreicher, hätte man auf dieser Stufe nicht eine dritte, quantitativ einschränkende, also auswählende Größe berücksichtigt: die Frequenz.

Die weitgehend automatisch generierte Liste enthält deshalb ‚nur' knapp 320.000 Stichwortkandidaten, die mit einer Mindestfrequenz von acht im Korpus vorkommen und in anderen Wörterbüchern enthalten sind. Die Frequenz von acht bedeutet, dass das betreffende Wort in seiner Grund- bzw. Nennform und/oder in den dazugehörigen Flexionsparadigmenformen mindestens achtmal im Korpus belegt ist. Die Zahl acht ist dabei zufällig jene Größe, mit deren Hilfe ein Schnitt bei der Extraktion aus dem Korpus vorgenommen werden konnte, durch den sich der Umfang der Stichwortkandidatenliste in idealer Weise beschränken ließ: Die Zahl der Stichwortkandidaten musste dabei die der (angestrebten) *elexiko*-Stichwörter deutlich übersteigen; die Differenz ist nötig, um im Zuge des manuell zu bewältigenden Teilschritts ausreichend Spielraum für die Streichung (beispielsweise fehlerhaft lemmatisierter) Stichwortkandidaten zu haben, ohne dabei Gefahr zu laufen, die Zahl der geplanten 300.000 *elexiko*-Stichwörter zu unterschreiten (was eine teilweise Wiederholung dieses Arbeitsschrittes zur Folge gehabt hätte).

## 2.2 Der ‚manuell' zu bewältigende Teilschritt

### 2.2.1 Die Indizierung der Stichwortkandidaten

Die schließlich in einer Datenbank zusammengefassten Stichwortkandidaten mussten in einem zweiten Arbeitsgang redaktionell dahingehend überprüft werden, inwieweit sie tatsächlich dem Status eines *elexiko*-Stichwortes entsprechen. Ein solcher Schritt war – wie Stichproben zeigten – in der Tat notwendig, um eindeutig fehlerhaft angesetzte Stichwortkandidaten auszufiltern sowie den Stichwortstatus aller übrigen (teil-) automatisch erzeugten Grundformen (darunter Varianten, mehr oder minder exotische Fremdwörter, Zitatwörter, Falschschreibungen, Namen u. a.) nochmals zu hinterfragen. Zu diesem Zweck wurde ein Lemmatisierungskonzept entwickelt, das hinsichtlich des Stichwortbegriffs auf die für *elexiko* relevanten lexikologisch-lexikografischen Fragestellungen zugeschnitten ist. Hierbei handelt es sich um Leitlinien, die Aufschluss darüber geben, welche Wörter in welcher Form als Stichwort anzusetzen sind und welche aus der Liste zu streichen sind (nähere Ausführungen dazu s. u.).

Alle ca. 320.000 Stichwortkandidaten wurden regelgeleitet und kompetenzgestützt überprüft sowie in die Kategorien ‚richtig', ‚unklar' und ‚falsch' eingeteilt und entsprechend in der Datenbank indiziert. Bei der überwiegenden Mehrheit der Stichwortkandidaten handelte es sich um Grundformen wie *laufen*, *lesen*, *Fahrrad*, *grün* etc., die mit einem ‚r' markiert wurden und damit als ‚richtiges' *elexiko*-Stichwort in der Datenbank ausgewiesen wurden.

Daneben fanden sich jedoch auch etliche Formen, die als fragwürdig, d. h. bis zu einer nochmaligen Korpusüberprüfung unklar, oder aber offenkundig als falsch eingestuft und entsprechend gekennzeichnet, also bis auf Weiteres aus der *elexiko*-Stichwortliste gestrichen wurden. Das Ursachenspektrum, das eine Ausmusterung notwendig machte, ist äußerst breit: So kann es sein, dass ‚falsche' Grundformen vom automatischen Lemmatisierer generiert worden sind; zur Illustration mag stellvertretend ein Beispiel genügen: *Metall-*, *Betonstrebe* sind Komposita zum Substantiv *die Strebe*. Außerdem gab es jedoch auch Stichwortkandidaten wie *Autonomiestrebe*, *Gewinnstrebe*, *Leistungsstrebe* etc., also eigentlich Zusammensetzungen mit dem substantivierten Infinitiv *das Streben*; hier jedoch hat der Algorithmus des Lemmatisierers das zum Stamm gehörige *-n* als Flexionsmorphem interpretiert (etwa Plural von *Strebe*), und infolgedessen in falscher Analogie abgetrennt. Das Beispiel der fehlerhaften Segmentierungen zeigt, wie allgemein gültige Subsumierungsregeln in wenigen Ausnahmefällen zu derartigen, z. T. kuriosen Wortneuschöpfungen führ-

ten, die natürlich nicht in *elexiko* gehören.[6] Es lässt zugleich die komplexen Mechanismen hinter dem automatischen Lemmatisieren erahnen. Zudem waren auch konzeptionell bedingte Ausschlüsse von Stichwortkandidaten vorzunehmen, beispielsweise bestimmter Flexionsformen eines Wortes (s. u.).

Knapp 300.000 schließlich mit ‚richtig' markierten Stichwortkandidaten, die auf diese Weise das Prädikat ‚*elexiko*-Stichwort'[7] erhielten, steht eine Menge von ca. 20.000 mit ‚unklar' bzw. ‚falsch' gekennzeichneten Einheiten gegenüber. Erstere wurden nach der Angleichung an die neue Rechtschreibung als *elexiko*-Stichwortliste Anfang 2004 ins Netz gestellt[8] und diese bilden damit das erste Nachschlageangebot für die BenutzerInnen als auch das Fundament für den weiteren Ausbau von *elexiko*; die Menge der vorläufig ausgemusterten Stichwortkandidaten bedarf einer weiteren Untersuchung, d. h. sie muss in einem vertiefenden zweiten Arbeitsgang subkategorisiert und ggf. auch nochmals im Korpus geprüft werden, um die Zweifelsfälle zu entscheiden; diese sind entweder nachträglich der *elexiko*-Stichwortliste zuzuführen oder auszuschließen[9], also in einer gesonderten Menge zusammenzufassen.

Von den technisch bedingten, meist schnell und eindeutig zu entscheidenden Zweifelsfällen sind grundsätzlich solche Wortformen zu unterscheiden, deren Status als Stichwort oder Einstufung als Nichtstichwort nur lexikografisch-konzeptionell begründbar ist. Hierunter fallen neben speziellen unregelmäßigen Formen des Flexionsparadigmas z. B. auch Pluraliatantumformen, Partizipien, Movierungen, Neubildungen und viele andere Sonderfälle, die allein im Kontext eines umfassenderen Lemmatisierungskonzepts zu betrachten sind. Standen bislang Vorgänge auf verfahrenstechnisch-praktischer Seite und deren stark vereinfachende Wiedergabe im Vordergrund, so verlagert sich nun das Augenmerk auf theoretisch-konzeptionelle Aspekte.

### 2.2.2 Stichwort – Lemmatisierungskonzept

Die skizzierte kompetenz- und regelbasierte Indizierung wurde vor dem Hintergrund folgender Leitfrage durchgeführt: Welche Wörter bzw. Wortformen sollen in welcher Schreibung den Status eines *elexiko*-Stichworts erhalten, oder anders formuliert: Welche Wörter werden BenutzerInnen (neben einer

---

6 *Achillesvers, Alphornblase, Augenreibe, Denkmalgeschütz, Drachenfliege, glitzerfummeln* und *Senftuba* gehören zu einer Reihe besonders ‚gelungener' Fehlansetzungen.

7 Nur knapp ein Drittel besaß zur Zeit der automatischen Erstellung der Stichwortkandidatenliste eine Frequenz von mehr als 50 Treffern im Korpus.

8 Siehe http://hypermedia.ids-mannheim.de/pls/elexiko/elexikoewl.ansicht.

9 Die Analyse der vom automatischen Lemmatisierer fehlerhaft generierten Stichwortkandidaten kann u. a. für dessen Modifizierung nutzbar gemacht werden.

Vielzahl anderer Informationen) später als Stichwort (vermutlich) suchen? Die Beantwortung dieser komplexen Fragestellung wird anschließend begleitet von Überlegungen zur Funktion des Stichwortes als abstrakter Wörterbucheinheit mit Blick auf das für *elexiko* entwickelte Lemmatisierungskonzept.

Die Begriffe ‚Stichwort' bzw. ‚Lemma' werden in den großen deutschen einsprachigen Wörterbüchern i. d. R. nicht definitorisch eingeführt, sondern gewissermaßen als aus der Praxis der Wörterbuchbenutzung heraus ‚traditionell bekannt' vorausgesetzt und entsprechend unkommentiert verwendet (vgl. z. B. Brockhaus-Wahrig, 9, Duden 2001, 23 ff., WDG 18 ff.). Aber auch beim Duden digital und auf den Seiten des DWDS sucht man vergeblich nach einer Stichwortbestimmung, sodass die Vermutung nahe liegt, hier werden die impliziten (und nebenbei: durchaus berechtigten) Annahmen der Printwörterbuchlexikografie einfach übernommen.

Wiegand hat sich der Aufgabe angenommen, die „meistens synonym verwendeten Termini Lemma und Stichwort" (Wiegand 1983, 403) zu präzisieren, allerdings unter der Einschränkung, „daß es nicht darum geht, Sprachregelungen für den lexikographischen Alltag in der Wörterbuchwerkstatt anzubieten", denn „in der lexikographischen „Werkstattsprache" ist der Gebrauch dieser Ausdrücke meist regelhaft" (ebd.).

Da im vorliegenden Beitrag aus der Werkstatt berichtet wird, die Perspektive von *elexiko* im Mittelpunkt steht, ist es also nicht nötig, auf alle Facetten dieser Theoriedebatte einzugehen (vgl. neben Wiegand 1983 hierzu u. a. auch Wolski 1989 sowie Reichmann 1989, bes. 37 ff.). In *elexiko* wurde der Begriff ‚Stichwort' bewusst dem des ‚Lemmas' vorgezogen und primär gemäß der spezifischen Anforderungen von *elexiko* entwickelt. Für diese Entscheidung spricht, dass die Bedeutung von ‚Lemma' im Werkstatt- und vor allem im wissenschaftlichen Diskurs oft unscharf ist (vgl. Wiegand 1983, 411 ff.); eine terminologische Präzisierung über den zweifelsfrei vorhandenen ‚common sense' hinaus schiene in jedem Falle angebracht. Dafür spricht zudem, dass die elektronische Publikationsform von *elexiko* andere Anforderungen an die Konzeption des ‚Stichworts' stellt als jene, die für das ‚Lemma' der traditionellen Printlexikografie formuliert wurden. Die Tatsache, dass in vielen retrodigitalisierten Wörterbüchern aber auch darüber hinaus vielfach eine funktional ausgerichtete Definition des Lemmas nicht explizit vorgenommen, diese vielmehr stillschweigend aus der (zugrunde gelegten) Printversion übernommen wird[10], bestätigt – pointiert formuliert – die Vermutung: Ein als digital bezeichnetes Wörterbuch ist so gesehen eine bloße Kopie des Papierwörterbuches auf eine Wörterbuch-CD.

---

[10] Entsprechende Stichproben basieren auf Brockhaus-Wahrig, Duden (2001), WDG sowie Duden digital und DWDS.

Ursachen, die eine kritische Hinterfragung des traditionellen Lemma-Begriffs vor dem Hintergrund neuer Publikationsformen zwingend nötig machen, sind zumindest die Folgenden:

- Die durch das Alphabet vorgegebene makrostrukturelle Ordnung eines Printwörterbuches kann (und sollte) zugunsten anderer Zugriffs-, Recherche- bzw. Navigationsstrukturen im digitalen bzw. Online-Wörterbuch aufgegeben werden, mit deren Hilfe sich NutzerInnen auf einer ganz anderen Ebene orientieren können. Das online-Stichwort bildet nur mehr eine von vielen möglichen Zugriffsstrukturen, derer sich BenutzerInnen bedienen können.
- Der die Printwörterbuchlexikografie prägende, durch den notorischen Platzmangel hervorgerufene Zwang zur Datenkomprimierung spielt in der Onlinepublikation (bedingt auch bei der Publikation auf digitalen Datenträgern) keine Rolle, sodass die lexikografische Textverdichtung weitgehend rückgängig gemacht werden kann. Auf das traditionelle Lemma übertragen bedeutet dies, dass sich die Frage, ob beispielsweise lexikografische Kommentarsymbole (etwa Silbentrennungsangaben, Homonymenindizes[11] u. a., vgl. Wiegand 1983, 437 ff., Wolski 1989, 362) zum ‚Stichwort' gehören oder nicht, gar nicht stellt. Vielmehr bleibt nur das Lemmazeichen (als Lemmazeichengestaltangabe) übrig, und zwar als

  > „dasjenige [Langue-] Zeichen, um das es in einem Wörterbuchartikel geht bzw. zu dessen Repräsentation das Lemma angesetzt wird, d. h.: das Lemmazeichen ist dasjenige Zeichen, das lexikographische bearbeitet ist" (Wiegand 1983, 428 f.).

  Die anderen erwähnten Informationseinheiten können, ja sollten im Zuge späterer Recheremöglichkeiten vom Stichwort losgelöst, und an anderer Stelle in der Datenstruktur verankert und eigens ausgezeichnet werden (zu Einzelheiten siehe C. Müller-Spitzer, Die Modellierung lexikografischer Daten, in diesem Band); Stichwort und Lemmazeichen sind also weitgehend identisch. Von der Printlexikografie her bekannte Unterarten des Lemmas wie Nest- oder Nischenlemma (vgl. u. a. Wiegand 1983, 432 ff.) müssen gleichfalls nicht unterschieden werden, da deren Funktion (platzsparendes Anordnen und umgekehrt Finden entsprechender Einheiten) bei der online-Recherche anders geregelt wird. An die Stelle der aus Wörterbüchern hinlänglich be-

[11] Die in der lexikografischen Praxis nicht einfach zu treffende Unterscheidung zwischen Polysemie und Homonymie wird in *elexiko* zugunsten eines alternativen Modells aufgegeben, siehe U. Haß, Das Bedeutungsspektrum, in diesem Band.

kannten Verweise vom Typ ***dritt-*** *siehe acht* können gedoppelte Informationen unter zwei Stichwörtern anführt werden.

- Die Definition des Stichwortes muss zumindest in zentralen Punkten mit den Annahmen der Zähleinheiten aus der automatischen Lemmatisierung übereinstimmen.[12]

Zur Veranschaulichung des Gesagten dient zunächst ein vereinfachtes Beispiel; daraus kann anschließend ein aus dem Arbeitsprozess heraus gewonnener Lösungsvorschlag abgeleitet werden.

In Printwörterbüchern findet man i. d. R. eine Auswahl an Flexionsformen lemmatisiert, etwa *sah* (zu *sehen*) *ist* (zu *sein*) *Indices* (zu Index) u. v. a. Derartige Flexionsformen wird man in der *elexiko*-Stichwortliste vergeblich suchen, denn: In *elexiko* haben Stichwörter einzig und allein eine Etikettierungsfunktion; das Lemmazeichen bildet gewissermaßen die Überschrift für einen Wortartikel. In der Printlexikografie hingegen besitzen Stichwörter bzw. Lemmata zusätzlich u. a. folgende Funktionen:

- Sie dienen im Rahmen der Makrostruktur der alphabetischen Sortierung also dem Suchen bzw. Finden von Informationen zu einem bestimmten Wort.
- Diesem Ziel, Informationen im Wörterbuch aufzufinden, dient auch der Verweis: So werden etwa (wie oben angedeutet) bestimmte unregelmäßige Flexionsformen häufig nur deshalb als Stichwort angesetzt, um den WörterbuchbenutzerInnen das Suchen und Finden zu erleichtern: Diese werden dann an der alphabetisch ‚richtigen' Stelle fündig und durch das Verweisstichwort von dort auf die dazu gehörende, jedoch alphabetisch entfernten Grundform samt dazugehörigem Wortartikel verwiesen, also von *sah* zu *sehen* usw. Derartige Verweise werden in *elexiko* über Links oder über das Recherchesystem geregelt: Für die BenutzerInnen soll später einmal z. B. bei Eingabe von *ging* ‚automatisch' der Wortartikel von *gehen* sichtbar werden mit der Information, um welche Form es sich bei *ging* im Verhältnis zu *gehen* handelt.

---

12 Dass hier andere Problemstellungen zu berücksichtigen sind, haben die knappen Ausführungen zum automatischen Lemmatisieren gezeigt; das unterstreichen auch zwei sprachstatistisch ausgerichtete Wörterbücher, nämlich das Frequenzwörterbuch zur deutschen Zeitungssprache sowie das Häufigkeitswörterbuch der gesprochenen Sprache, wobei sich Rosengren im Gegensatz zu Ruoff (1981) auch intensiv theoretisch mit der Definition der Zähleinheit auseinandersetzt (vgl. Rosengren 1969 und Rosengren 1972-1977, bes. Bd. 2, IX ff.).

Im Lemmatisierungskonzept von *elexiko* ist deshalb festgehalten, dass Flexionsformen nicht den Status als *elexiko*-Stichwort erhalten, sie wurden deshalb aus der Stichwortkandidatenliste gestrichen. Als reine Suchwörter sind sie im Kontext des Informationssystems als Ganzes funktional auf einer völlig anderen Ebene angesiedelt. Das bedeutet ferner: Bestimmte Formen des Flexionsparadigmas werden im Zuge der Wortartikelproduktion als eigens ausgezeichnete Wortformen einer speziellen Liste zugeführt, und somit dem Bereich der Recherche- und Zugriffsstrukturen beigeordnet. Mit anderen Worten: In *elexiko* gibt es:

- Stichwörter (die als Lemmazeichen regelgeleitet in der Grundform und orthografisch normgerecht angesetzt werden, s. u.) auf der Ebene der Wortartikelproduktion und
- Suchwörter (darunter fallen die traditionellen Verweislemmata) auf der Ebene der Recherchestrukturen: Alle Stichwörter sind selbstverständlich zugleich Suchwörter, hinzu kommen bestimmte Formen aus dem Flexionsparadigma sowie orthografische und morphologische Varianten.

Ausnahmen, die im Rahmen der Lemmakonzeption eigens behandelt werden müssen, sind Fälle, bei denen etwa die eigentliche Grundform nicht gebildet werden kann oder üblicherweise nicht gebildet wird bzw. diese im Korpus nicht gefunden wurde, also etwa Pluraliatantum- und bestimmte Partizip-Formen, z. B. *Spinnenarme*, *Stadtväter*, *Flitterwochen*, *Fidschiinseln*; *begabt*, *aufgedunsen*.

Sofern Flexionsformen die Voraussetzung der semantisch-grammatischen Eigenständigkeit erfüllen, werden sie als eigenes Stichwort angesetzt. Verallgemeinert lässt sich daraus folgender Leitgedanke für die Stichwortansetzung in *elexiko* ableiten: Oberstes Gebot bei der Beantwortung der Frage, was primär als Stichwort anzusetzen ist, ist die Eigenständigkeit, d. h. neben den Grundformen i. e. S. werden in die *elexiko*-Stichwortliste nur solche Wortformen aufgenommen, die sich inhaltlich und/oder formal nicht über die Grundform ‚erschließen' und beschreiben lassen, bzw. für die eine Grundform im Korpus nicht (oder nur unzureichend) belegt ist. Im Einzelfall wird sich das erst im Kontext der Wortartikelproduktion entscheiden lassen; ebenso, ob es nötig sein wird, sekundär bestimmte Flexionsformen wie Plurale, (suppletive) Steigerungsformen etc. als eigene Stichwörter anzusetzen, weil sie als grammatisch/semantisch eigenständig einzustufen sind: Die Frage des Status von bestimmten grammatischen Formen wird folglich auch erst beim Schreiben der Wortartikel zu beantworten sein. Solange noch kein entsprechender Wortartikel verfasst wurde, besitzt jedes *elexiko*-Stichwort deshalb in gewisser Weise

hypothetischen, abstrakten Charakter: Es handelt sich also nicht um ein statisches, sondern um ein dynamisches Stichwortkonzept. Dass neben dem zentralen Kriterium der Eigenständigkeit auch noch andere Faktoren die Stichwortansetzung beeinflussen können, kann am Beispiel der Frequenz gezeigt werden: Sie spielt etwa bei der Entscheidung eine Rolle, was als Stichwort bzw. was als morphologisch Variante (*mickrig* oder *mickerig*) anzusetzen ist (siehe dazu A. Klosa, Orthografie und morphologische Varianten, in diesem Band).

Das Skizzierte ist die Grundlage für eine Reihe von Begriffsdefinitionen, wie sie im Zusammenhang mit der Stichwortansetzung speziell für *elexiko* entwickelt wurden (und fortlaufend weiterentwickelt werden sollen):

- ***elexiko*-Stichwortliste**: Liste aller in *elexiko* gemäß den Lemmatisierungsprinzipien angesetzten Stichwörter, der in *elexiko* zu beschreibende Wortbestand (vgl. http://hypermedia.ids-mannheim.de/pls/elexiko/elexiko_ewl.ansicht)
- **Grundform (auch Nennform, Zitierform, kanonische Form)**: Unflektierte Form eines Wortes, z. B. Infinitiv (bei Verben), Nom., Sing. (bei Nomen), Positiv (bei Adjektiven) in orthografisch normgerechter Schreibung.
- **Lemma** wurde u. a. wegen seiner Vieldeutigkeit bzw. Implikationen (seitens der Printlexikografie) in *elexiko* ersetzt durch Stichwort.
- **Lemmatisierung** bezieht sich auf die automatische Grundformrückführung durch den Lemmatisierer einerseits, andererseits auf die kriteriengeleitete Festsetzung der Stichwörter. Die Schritte von der automatischen Lemmatisierung zur Stichwortkandidatenliste, von dieser mittels kriteriengeleiteter Kompetenzprüfung hin zur *elexiko*-Stichwortliste lassen sich als primäre Lemmatisierung bezeichenen. Darüber hinaus wird im Zuge der Wortartikelproduktion eine Reihe von Wörtern bzw. Wortformen, die sich als eigenständig erweisen, nachlemmatisiert werden müssen; dies kann als sekundäre Lemmatisierung verstanden werden.
- **Stichwort**: Redaktionell-kriteriengestützt geprüfte und orthografisch normgerechte (abstrahierte) Stammform, die in Form der Lemmazeichengestaltangabe realisiert wird und deren Funktion die Etikettierung des *elexiko*-Wortartikels ist. Neben den Einwortstichwörtern, die hier im Zentrum der Ausführungen stehen, ist geplant, die *elexiko*-Stichwortliste durch (methodisch auf anderen Wegen zu ermittelnde) Mehrworteinheiten und Wortbildungsmorpheme zu ergänzen.
- **Stichwortkandidat**: Vom automatischen Lemmatisierer erzeugte Grundform eines Wortes unter der die unterschiedlichen Wortformen

des Flexionsparadigmas subsumiert werden. Erfüllt der Stichwortkandidat im Zuge der regelgeleiteten, kompetenzgestützten Prüfung die einschlägigen Kriterien, erhält er den Status eines Stichwortes.

- **Stichwortkandidatenliste**: Auf Korpusbasis vom automatischen Lemmatisierer erzeugte, halbautomatisch durch Abgleich mit Listen anderer Wörterbücher bereinigte Liste der Stichwortkandidaten, auf deren Basis die *elexiko*-Stichwortliste erarbeitet wurde.
- **Suchwort**: Alle *elexiko*-Stichwörter sowie bestimmte Paradigmaformen, die BenutzerInnen verwenden können, um auf der Ebene der Zugriffsstrukturen in *elexiko* nach Wortartikeln zu suchen.
- **Wortform**: flektierte Form eines Wortes im Gegensatz zur Grundform.

Damit kommt *elexiko* einer Forderung Wolskis nach:

> Der wissenschaftliche Gegenstand ‚Lemma' [bzw. Stichwort, U. Sch.] muß [...] im Rahmen einer Theorie der lexikographischen Sprachbeschreibung so differenziert erfaßt werden, wie es aufgrund der in Wörterbüchern vorhandenen Gestaltungsvielfalt erforderlich ist; gegebenenfalls können als zu weitgehend erkannte Differenzierungen in einem späteren Schritt wieder zurückgenommen werden (schließlich geht es in keiner Wissenschaft und in keiner wissenschaftlichen Praxis um die Vermittlung ewiger Wahrheiten). (Wolski 1989, 369)

Das gilt natürlich in besonderem Maße auch für die konkreten Lemmatisierungskriterien von *elexiko*; sie sollen (in Ausschnitten) die Überlegungen abrunden (siehe zu Details über Orthografie des Stichworts [= Lemmazeichengestaltangabe] und dessen morphologische Varianten in *elexiko* A. Klosa, Orthografie und morphologische Varianten, in diesem Band).

Prinzipiell ist es sinnvoll, für die Entwicklung von Kriterien zur Stichwortansetzung nicht nur zwischen Wortarten, sondern auch zwischen offenen und geschlossenen Wortklassen zu unterscheiden. Für die potentiell unendliche Zahl von anzusetzenden Stichwörtern aus den offenen Wortklassen sind Richtlinien erstrebenswert, die möglichst generelle Gültigkeit besitzen, d. h. sie sollten für die überwiegende Mehrheit der Fälle angewandt werden können, und möglichst wenige Ausnahmen zur Konsequenz haben.[13] Für Adjektive, Nomen und Verben gilt Folgendes:

---

[13] Demgegenüber können bei den Richtlinien für die (relativ) geschlossenen Klassen Sonderwege eingeschlagen werden, die den speziellen Anforderungen dieser Klassen angemessen sind; diese werden daher hier ausgeklammert.

2.2.2.1 Wortart: Adjektiv

a) genereller Stichwortansatz

Positiv; Komparativ bzw. Superlativ und auch Suppletivformen werden nur dann als Stichwort angesetzt, wenn sie semantisch eigenständig sind; dies wird sich im Einzelfall erst bei der Wortartikelproduktion entscheiden lassen. I. d. R. sind es Suchwörter.

b) dynamischer Stichwortansatz (d. h. Statusänderung ist unter bestimmten Bedingungen möglich) und Sonderfälle

- (unregelmäßige) Komparative wie *höher* zu *hoch*, *jünger* als absoluter Komparativ zu *jung* gelten als Suchwort.
- unregelmäßige Superlative wie *jüngste* zu *jung* gelten als Suchwort.
- Adjektive vom Typ *blamabel – blamable(r/s)*: Die Form *blamabel* ist Stichwort, alle Flexionsformen sind Suchwörter. Ordinalzahlen werden in der attributiv gebrauchten Form aufgenommen, z. B. *achte*.
- *Brüsseler* als Einwohnerbezeichnung und *Brüsseler* als indeklinables Adjektiv werden auf auf Lesartenebene unterschieden.

2.2.2.2 Wortart: Nomen

a) genereller Stichwortansatz

Nominativ Singular; wenn dieser, wie z. B. bei einem Pluraliatantum nicht existiert: Nominativ Plural.

b) dynamischer Stichwortansatz (d. h. Statusänderung ist unter bestimmten Bedingungen möglich) und Sonderfälle

- unregelmäßige Pluralformen (z. B. *Indices*, *Indexe*) gelten als Suchwort.
- Im Falle der Klassifikation als Pluraliatantum muss von Fall zu Fall während der Wortartikelproduktion entschieden werden, wobei das Verhältnis von singularischen und pluralischen Korpusbelegen zu berücksichtigen ist.
- Diminutiva auf *-chen* und *-lein* werden, sofern sie in der *elexiko*-Stichwortliste enthalten sind, zunächst in der Stichwortliste belassen.

- Movierungen (*Beschauerin* zu *Beschauer*) werden prinzipiell als eigenes Stichwort behandelt.
- Substantive vom Typ *Angestellte(r)* werden (wie mit unbestimmtem Artikel) sowohl in der maskulinen (*Angestellter*) als auch in der femininen (*Angestellte*) Form aufgenommen.
- Alle usuellen, in der Stichwortkandidatenwortliste enthaltenen Komposita werden zu *elexiko*-Stichwörtern.[14] Zu bedenken ist, dass der automatische Lemmatisierer Bindestrichkomposita i. d. R. segmentiert, sodass evtl. eine Reihe von Schreibvarianten unterschlagen wurde; falls man auf diese im Zuge der Wortartikelproduktion aufmerksam wird, ist ein nachträglicher Ansatz als Variante erforderlich.
- Alle usuellen, in der Stichwortkandidatenliste enthaltenen Fremdwörter werden zu *elexiko*-Stichwörtern. Sie werden in der assimilierten deutschen Schreibung angegeben (z. B. *Perestroika*; *Perestrojka* soll dagegen Variante und Suchwort werden), nicht berücksichtigt werden Zitatwörter (z. B. aus englischsprachigen Kinofilmtiteln: *Man | in | Black*)
- Alle usuellen, in der Stichwortkandidatenliste enthaltenen substantivierten Infinitive, Adjektive usw. werden zu *elexiko*-Stichwörtern (grammatisch/semantische Selbstständigkeit vorausgesetzt).
- Alle usuellen, in der Stichwortkandidatenliste enthaltenen Namen werden zu *elexiko*-Stichwörtern; obwohl diese bislang vielfach von der lexikografischen Beschreibung ausgeschlossen wurden, spricht nicht nur das Prinzip der Korpusbasiertheit für eine Aufnahme. Die Richtlinie ist allein deshalb gerechtfertigt, weil sehr oft eine Unterscheidung zwischen Gattungsname und Eigenname erst im Zuge der Wortartikelproduktion vorgenommen werden kann, die *elexiko*-Stichwortliste also von vornherein eine Menge ‚versteckter' Namen enthält (z. B. *Hirsch*). Nur wenn Namen auch als Stichwörter angesetzt sind, ist es ferner möglich, bei der Korpusrecherche zu deren Bearbeitung auf eventuelle Appellativierungen zu stoßen, wie sie z. B. bei *Tesa* (‚Klebefilm'), *Uhu* (‚Klebstoff') usw. stattgefunden haben. Die Beschreibung von Namen ist grundsätzlich sinnvoll, da somit vielfältige Besonderheiten, v. a. in Bezug auf Orthografie, Grammatik und Wortbildung erfasst und diese NutzerInnen nicht vorenthalten werden, z. B. *Schweden*, aber: *der Irak*, *brandenburger Stadtverwaltung*, aber: *Brandenburger Tor* und *Brandenburgische Konzerte* (von Bach). Nicht als *elexiko*-Stichwort angesetzt wurden

[14] Usuell bedeutet in diesem Zusammenhang, dass das Stichwort mit einer Mindestfrequenz von 8 in den Korpora belegt und/oder durch den Wörterbuchabgleich gesichert sein muss.

vorläufig Namen von ‚Allerweltsstraßen' und ‚-plätzen' des Typs *Hofackerweg*; daneben gibt es natürlich auch solche wie *Weinstraße*, *Bergstraße*, die eigentlich nicht ausgeschlossen werden dürften. Auf diesem Feld muss noch systematisch nachgearbeitet werden.

- Abkürzungen werden (zunächst) nicht regelhaft berücksichtigt, und könnten später einen eigenen, komplett abarbeitbaren Wortschatzausschnitt (Modul) bilden; zu einer Ausnahme kann es kommen, wenn sich bei der Wortartikelbearbeitung für ein ‚Langwort' herausstellt, dass eine korrespondierende Abkürzung häufiger ist, als das entsprechende Langwort selbst: Für die Abkürzung wird dann gleichfalls ein Wortartikel erstellt.
- (Noch) nicht berücksichtigt werden ferner Siglen, fachsprachliche Zeiche (z. B. Au = Aurum, ‚Gold', fl = Flöte, f = forte, A = Kraftfahrzeugkennzeichen A für Augsburg, ha = Hektar) sowie für Zeichen wie @, ®, ☺, ;-), % etc.

2.2.2.3 Wortart: Verb

a) genereller Stichwortansatz

Infinitiv

b) dynamischer Stichwortansatz (d. h. Statusänderung ist unter bestimmten Bedingungen möglich) und Sonderfälle

- bestimmte Flexionsformen bei Verben wie *quoll*, *quölle* (zu) *quellen*; *muss* (zu *müssen*), *lädst*, *lädt* (zu *laden*), *nimmst*, *nimmt* (zu *nehmen*) gelten als Suchwort, d. h. im Rahmen der Grammatik angegebene (unregelmäßige) Flexionsformen werden gleichzeitig als Suchwörter ausgezeichnet.
- Partizipien werden zunächst primär als Stichwort angesetzt und als solche gelöscht, wenn sie grammatisch/semantisch nicht eigenständig sind, d. h. unter dem Infinitiv lexikografisch erschlossen werden können (siehe hierzu S. Erb, Partizipien, in diesem Band).
- Präverbfügungen (*danksagen*), die nach der Orthografiereform auch eine getrennt geschriebene Variante (*Dank sagen*) besitzen, werden unabhängig von ihrer (alten/neuen) Schreibung als Einworteinheit behandelt: Mit ausschlaggebend für diese Entscheidung ist die Festigkeit der Verbindung; darüber hinaus ist es auch eine Frage der Stimmigkeit, funktionsidentische Elemente im gleichen Beschreibungs-

rahmen zu behandeln. (Funktionsverbgefüge sind demgegenüber Mehrworteinheiten, u. a. sprechen die Kriterien der Reihenbildung und Varianz gegen eine Einstufung als Einworteinheit).

## 3. Zusammenfassung

Grafisch lässt sich die Entwicklung der *elexiko*-Stichwortliste wie folgt darstellen:

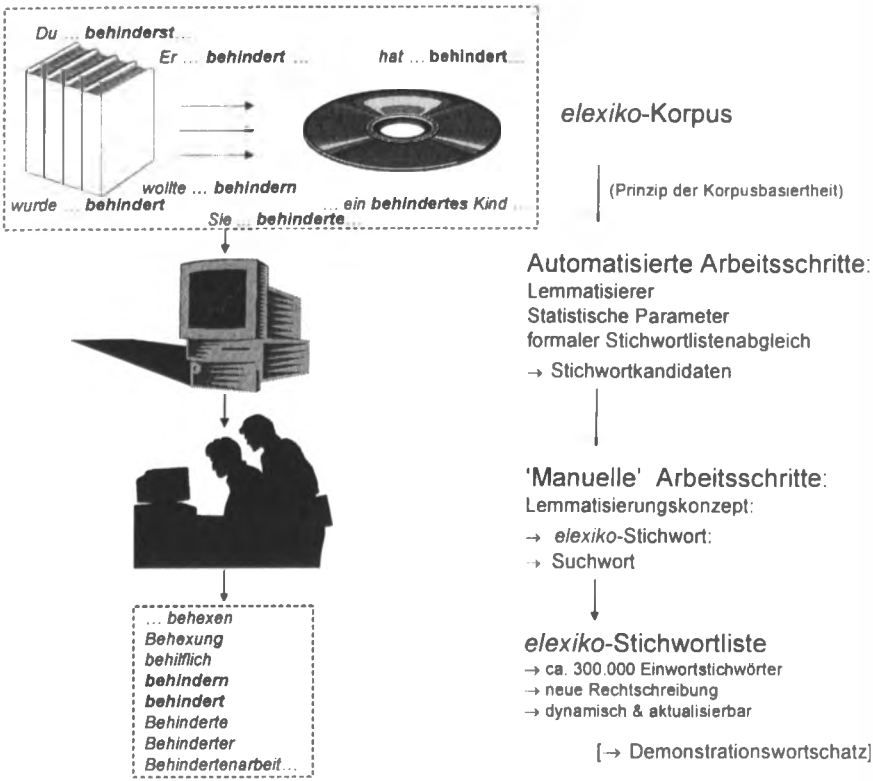


Abbildung 1: Die Erstellung der *elexiko*-Stichwortliste

Die zentralen Punkte sind:

1. Die *elexiko*-Stichwortliste enthält nach dem Indizierungsdurchgang ca. 300.000 Stichwörter; sie erhielten diesen Status im Rahmen des *elexiko*-Lemmatisierungskonzepts – zumindest bis zum Beweis des

Gegenteils: In Einzelfällen dürfte sich nämlich erst im Zuge der korpusbasierten Wortartikelproduktion verbindlich zeigen, ob zu jedem Stichwort ein eigener Wortartikel angelegt werden muss. Solange noch kein entsprechender Wortartikel verfasst wurde, besitzt das *elexiko*-Stichwort deshalb in gewissem Umfang hypothetischen Charakter. Es handelt sich daher nicht um ein statisches, sondern um ein dynamisches Konzept. So muss die *elexiko*-Stichwortliste beispielsweise auch aus Gründen der von einem online-Wörter-„Buch" zu Recht geforderten Aktualität die Voraussetzung der Veränderbarkeit erfüllen, also auch offen für neue Stichwörter sein.

2. Die Liste enthält auch Sonderwortschätze wie Fremdwörter, Fachwörter, Eigennamen, Dialektausdrücke etc., sofern sie ausreichend im *elexiko*-Korpus belegt sind. Die Mehrzahl der Abkürzungen wurde zunächst in einer gesonderten Menge erfasst und noch nicht berücksichtigt. (Dieser Bereich ist besonders anfällig für Segmentierungsprobleme beim automatischen Lemmatisieren, sodass hier u. U. das methodische Vorgehen modifiziert werden muss.
3. Die Liste enthält ferner keine Wortbildungs- bzw. Wortelementeinheiten und noch keine Mehrworteinheiten; diese beiden Wortschatzgruppen sind mit der skizzierten Vorgehensweise praktisch nicht zu ermitteln; dafür sind andere methodische Ansätze zu erarbeiten und anzuwenden.

## 4. Literaturverzeichnis

### 4.1 Wörterbücher

Brockhaus-Wahrig (1980) = Brockhaus-Wahrig. Deutsches Wörterbuch in sechs Bänden. Herausgegeben von Gerhard Wahrig u. a. Wiesbaden/Stuttgart.

Duden (2001) = Duden – Das große Wörterbuch der deutschen Sprache in zehn Bänden. Herausgegeben von der Dudenredaktion. Mannheim u. a.

Duden digital (2000) = Duden – Das große Wörterbuch der deutschen Sprache. PC-Bibliothek Version 2.01. Redaktionelle Bearbeitung: Werner Scholze-Stubenrecht, Dieter Mang u. a. Mannheim.

Rosengren, Inger (1972-1977): Ein Frequenzwörterbuch der deutschen Zeitungssprache. Die Welt. Süddeutsche Zeitung, 2 Bde. Lund.

Arno Ruoff (1981): Häufigkeitswörterbuch der gesprochenen Sprache: alphabetisch, rückläufig alphabetisch und nach Häufigkeit geordnet. Tübingen. (Idiomatica 8).

WDG (1980) = Wörterbuch der deutschen Gegenwartssprache. Herausgegeben von Ruth Klappenbach und Wolfgang Steinitz. Berlin 1980.

## 4.2 Forschungsliteratur

Rosengren, Inger (1969): Wort und Wortform. In: Studia Linguistica, 23, S. 103-113.

Reichmann, Oskar (1989): Lexikographische Einleitung. In: Anderson, Robert R./Goebel, Ulrich/Reichmann, Oskar (Hg.) (1989): Frühneuhochdeutsches Wörterbuch. Band 1. Berlin/New York. S. 10-164.

Wiegand, Herbert-Ernst (1983): Was ist eigentlich ein Lemma? Ein Beitrag zur Theorie der lexikographischen Sprachbeschreibung. In: Wiegand, Herbert-Ernst (Hg.) (1983): Studien zur neuhochdeutschen Lexikographie III. Hildesheim. S. 401-474 (Germanistische Linguistik 1-4/82).

Wiegand, Herbert-Ernst (1989): Aspekte der Makrostruktur im allgemeinen einsprachigen Wörterbuch: alphabetische Anordnungsformen und ihre Probleme. In: Hausmann, Franz Josef/Reichmann, Oskar/Wiegand, Herbert Ernst/Zgusta, Ladislav (Hg.) (1989): Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch der Lexikographie. Erster Teilband. Berlin/New York. S. 371-409 (Handbücher zur Sprach- und Kommunikationswissenschaft 5.1).

Wolski, Werner (1989): Das Lemma und die verschiedenen Lemmatypen. In: Hausmann, Franz Josef/Reichmann, Oskar/Wiegand, Herbert Ernst/ Zgusta, Ladislav (Hg.) (1989): Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch der Lexikographie. Erster Teilband Berlin/New York. S. 360-371. (Handbücher zur Sprach- und Kommunikationswissenschaft 5.1).

## 4.3 Internetressourcen

DWDS = Das Digitale Wörterbuch der deutschen Sprache des 20 Jahrhunderts: http://www.dwds.de/ (letzter Zugang September 2004).

*elexiko*: http://www.elexiko.de (letzter Zugang September 2004).

IDS-Korpopa geschriebener Gegenwartssprache: http://www.ids-mannheim.de/kt/projekte/korpora/.

*elexiko*-Stichwortliste: http://hypermedia.ids-mannheim.de/pls/elexiko/elexiko ewl.ansicht (letzter Zugang September 2004).

Lemmatisierer: http://www.ids-mannheim.de/kt/dokumente/glemmrep.pdf (letzter Zugang September 2004).